PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G06F 17/30

A2

(11) International Publication Number: WO 99/36863

(43) International Publication Date: 22 July 1999 (22.07.99)

(21) International Application Number: PCT/IB99/00033

(22) International Filing Date: 13 January 1999 (13.01.99)

(30) Priority Data:
09/006,657 13 January 1998 (13.01.98) US

(71) Applicant: KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(71) Applicant *(for SE only)*: PHILIPS AB [SE/SE]; Kottbygatan 7, Kista, S-164 85 Stockholm (SE).

(72) Inventor: DIMITROVA, Nevenka; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

(74) Agent: FAESSEN, Louis, M., H.; Prof. Holstlaan 6, NL-5656 AA Eindhoven (NL).

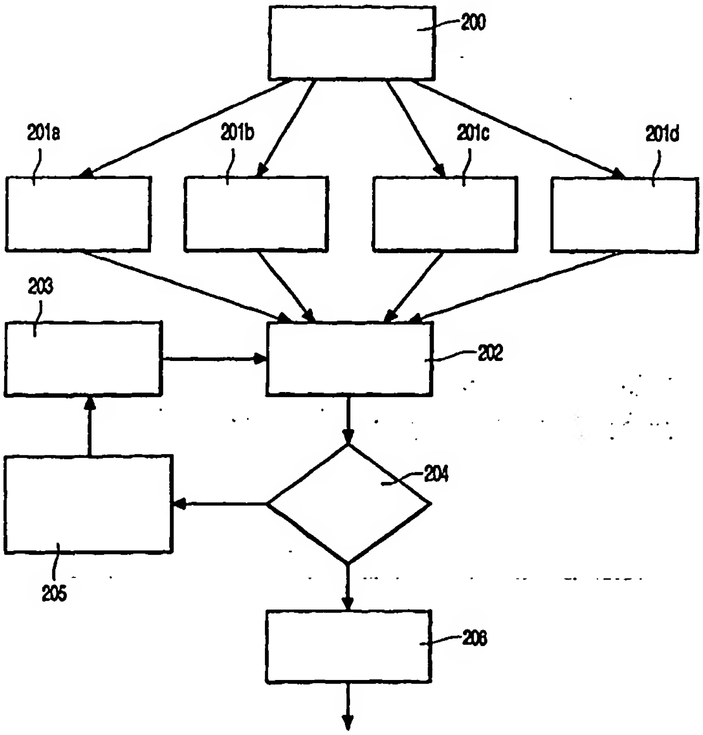
(81) Designated States: JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: MULTIMEDIA COMPUTER SYSTEM WITH STORY SEGMENTATION CAPABILITY AND OPERATING PROGRAM THEREFOR

(57) Abstract

Information is generated to support selective retrieval of a video sequence. This involves providing a set of models, each for recognizing a sequence of symbols. The symbols include symbols that represent key frames, audio and text properties associated with segments of the video sequence. A matching model is selected, which allows recognition of a sequence of symbols that are coupled to successive segments of the video sequence so that the key frame and audio and/or text properties satisfy the selected matching model. A reference to the matching model is used as a selection criterion for retrieving the video sequence. Optionally, a new model is constructed when no matching model for the video sequence is present in the set of models. The new model is constructed so that it allows recognition of the symbols of the video sequence. The new model is then used as selection criterion for retrieving the video sequence.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

Multimedia computer system with story segmentation capability and operating program therefor.

BACKGROUND OF THE INVENTION

The present invention relates generally to multimedia systems, including hybrid television-computer systems. More specifically, the present invention relates to story segmentation systems and corresponding processing software for separating an input video signal into discrete story segments. Advantageously, the multimedia system implements a finite automaton parser for video story segmentation.

Popular literature is replete with images of personal information systems where the user can merely input several keywords and the system will save any news broadcast, either radio or television broadcast, for later playback. To date, only computer systems running news retrieval software have come anywhere close to realizing the dream of a personal news retrieval system. In these systems, which generally run dedicated software, and may require specialized hardware, the computer monitors an information source and downloads articles of interest. For example, several programs can be used to monitor the Internet and download articles of interest in background for later replay by the user. Although these articles may include links to audio or video clips which can be downloaded while the article is being examined, the articles are selected based on keywords in the text. However, many sources of information, e.g., broadcast and cable television signals, cannot be retrieved in this manner.

The first hurdle which must be overcome in producing a multimedia computer system and corresponding operating method capable of video story segmentation is in designing a software or hardware system capable of parsing an incoming video signal, where the term video signal denotes, e.g., a broadcast television signal including video shots and corresponding audio segments. For example, U.S. Patent No. 5,635,982 discloses an automatic video content parser for parsing video shots so that they can be represented in their native media and retrieved based on their visual content. Moreover, this patent discloses methods for temporal segmentation of video sequences into individual camera shots using a twincomparison method, which method is capable of detecting both camera shots implemented by sharp break and gradual transitions implemented by special editing techniques, including dissolve, wipe, fade-in and fade-out; and content-based keyframe selection of individual shots by analyzing the temporal variation of video content and selecting a key frame once the difference of content between the current frame and a preceding selected keyframe exceeds a set of preselected thresholds. The patent admits that such parsing is a necessary first step in any video indexing process. However, while the automatic video parser is capable of parsing a received video stream into a number of separate video shots, i.e., cut detection, the automatic video processor is incapable of video indexing the incoming video signal based on the parsed video segments, i.e., content parsing.

While there has been significant previous research in parsing and interpreting spoken and written natural languages, e.g., English, French, etc., the advent of new interactive devices has motivated the extension of traditional lines of research. There has been significant investigation into processing isolated media, especially speech and natural language and, to a lesser degree, handwriting. Other research has focused on parsing equations (e.g., a handwritten "5+3"), drawings (e.g., flow charts), and even face recognition, e.g., lip, eye, and head movements. While parsing and analyzing multimedia presents an even greater challenges with a potentially commensurate reward, the literature is only now suggesting the analysis of multiple types of media for the purpose of resolving ambiguities in one of the media types. For example, the addition of a visual channel to a speech recognizer could provide further visual information, e.g., lip movements, and body posture, which could be used to help in resolving ambiguous speech. However, these investigations have not considered using the output of, for example, a language parser to identify keywords which can be associated with video segments to further identify these video segments.

The article by Deborah Swanberg eta al. entitled "Knowledge Guided Parsing in Video Databases" summarized the problem as follows:
"Visual information systems require both database and vision system capabilities, but a gap exists between these two systems: databases do not provide image segmentation, and vision systems do not provide database query capabilities. . . : The data acquisition in typical alphanumeric databases relies primarily on the user to type in the data. Similarly, past visual databases have provided keyword descriptions of the visual descriptions of the visual data, so

data entry did not vary much from the original alphanumeric systems. In many cases, however, these old visual systems did not provide a sufficient description of the content of the data."

The paper proposed a new set of tools which could be used to: semiautomatically segment the video data into domain objects; process the video segments to extract features from the video frames; represent desired domains as models; and compare the extracted features and domain objects with the representative models. The article suggests the representation of episodes with finite automatons, where the alphabet consists of the possible shots making up the continuous video stream and where the states contain a list arcs, i.e., a pointer to a shot model and a pointer to the next state.

In contrast, the article by M. Yeung et al., entitled "Video Content Characterization and Compaction for Digital Library Applications" describes content characterization by a two step process of labeling, i.e., assigning shots that are visually similar and temporally close to each other the same label, and model identification in terms of the resulting label sequence. Three fundamental models are proposed: dialogues, action; and story unit models. Each of these models has a corresponding recognition algorithm.

The second hurdle which must be overcome in producing a multimedia computer system and corresponding operating method capable of video story segmentation is in integrating other software, including text parsing and analysis software and voice recognition software, into a software and/or hardware system capable of content analysis of any audio and text, e.g., closed captions, in an incoming multimedia signal, e.g., a broadcast video signal. The final hurdle which must be overcome in producing a multimedia computer system and corresponding operating method capable of story segmentation is in designing a software or hardware system capable integrating the outputs of the various parsing modules or devices into a structure permitting replay of only the story segments in the incoming video signal which are of interest to the user.

What is needed is a multimedia system and corresponding operating program for story segmentation based on plural portions of a multimedia signal, e.g., a broadcast video signal. Moreover, what is needed is an improved multimedia signal parser which either effectively matches story segment patterns with predefined story patterns or which generates a new story pattern in the event that a match cannot be found. Furthermore, a multimedia computer system and corresponding operating program which can extract usable information from all of the included information types, e.g., video, audio, and text, included in a multimedia signal would be extremely desirable, particularly when the multimedia source is a broadcast television signal, irrespective of its transmission method.

SUMMARY OF THE INVENTION

Based on the above and foregoing, it can be appreciated that there presently exists a need in the art for a multimedia computer system and corresponding operating method which overcomes the above-described deficiencies. The present invention was motivated by a desire to overcome the drawbacks and shortcomings of the presently available technology, and thereby fulfill this need in the art.

The present invention is a multimedia computer system and corresponding operating method capable of performing video story segmentation on an incoming multimedia signal. According to one aspect of the present invention, the video segmentation method advantageously can be performed automatically or under direct control of the user.

One object of the present invention is to provide a multimedia computer system for processing and retrieving video information of interest based on information extracted from video signals, audio signals, and text constituting a multimedia signal.

Another object according to the present invention is to produce a method for analyzing and processing multimedia signals for later recovery. Preferably, the method generates a finite automaton (FA) modeling the format of the received multimedia signal. Advantageously, key words extracted from a closed caption insert are associated with each node of the FA. Moreover, the FA can be expanded to include nodes representing music and conversation.

Still another object according to the present invention is to provide a method for recovering a multimedia signal selected by the user based on the FA class and FA characteristics.

Yet another object according to the present invention is to provide a storage media for storing program modules for converting a general purpose multimedia computer system into a specialized multimedia computer system for processing and recovering multimedia signals in accordance with finite automatons. The storage media advantageously can be a memory device such as a magnetic storage device, an optical storage device or a magneto-optical storage device.

These and other objects, features and advantages according to the present invention are provided by a method of generating information to support selective retrieval of a video sequence, the method comprising
providing a set of models, each for recognizing a sequence of symbols;
selecting a matching model, which allows recognition of a sequence of symbols that are coupled to successive segments of the video sequence, the symbols including symbols that represent keyframes having properties prescribed by the model;
using a reference to the matching model as a selection criterion for retrieving the video sequence;
characterized in that the video sequence is temporally associated with at least one of audio information and text information, the symbols including symbols that represent at least one of audio and text properties associated with the segments in addition to the symbols that represent properties of the key frames, the matching model being selected so that the segments a sequence of symbols representing key frame and audio and/or text properties is recognized.

In an embodiment the method includes
constructing a new model, which allows recognition of the symbols of the video sequence;
adding said new model to the set of models when no matching model for the video sequence is present in the set of models;
using the new model as selection criterion.

Another aspect of the invention provides for by a storage medium for storing computer readable instructions for permitting a multimedia computer system receiving a multimedia signal containing unknown information, the multimedia signal including a video signal, an audio signal and text, to perform a parsing process on the multimedia signal to thereby generate a finite automaton (FA) model and to one of store and discard an identifier associated with the FA model based on agreement between user-selected keywords and keywords associated with each node of the FA model extracted by the parsing process. According to one aspect of the invention, the storage medium comprises a rewritable compact disc (CD-RW) and wherein the multimedia signal is a broadcast television signal.

These and other objects, features and advantages according to the present invention are provided by a storage medium for storing computer readable instructions for permitting a multimedia computer system to retrieve a selected multimedia signal from a plurality of stored multimedia signals by identifying a finite automaton (FA) model having a substantial similarity to the selected multimedia signal and by comparing FA characteristics associated with the nodes of the FA model with user-specified characteristics. According to one aspect of the present invention, the storage medium comprises a hard disk drive while the multimedia signals are stored on a digital versatile disc (DVD).

These and other objects, features and advantages according to the present invention are provided by a multimedia signal parsing method for operating a multimedia computer system receiving a multimedia signal including a video shot sequence, an audio signal and text information to permit story segmentation of the multimedia signal into discrete stories, each of which has associated therewith a final finite automaton (FA) model and keywords, at least one of which is associated with a respective node of the FA model. Preferably, the method includes steps for:

(a) analyzing the video portion of the received multimedia signal to identify keyframes therein to thereby generate identified keyframes;

(b) comparing the identified keyframes within the video shot sequence with predetermined FA characteristics to identify a pattern of appearance within the video shot sequence;

(c) constructing a finite automaton (FA) model describing the appearance of the video shot sequence to thereby generate a constructed FA model;

(d) coupling neighboring video shots or similar shots with the identified keyframes when the neighboring video shots are apparently related to a story represented by the identified keyframes;

(e) extracting the keywords from the text information and storing the keywords at locations associated with each node of the constructed FA model;

(f) analyzing and segmenting the audio signal in the multimedia signal into identified speaker segments, music segments, and silent segments

(g) attaching the identified speaker segments, music segments, laughter segments, and silent segments to the constructed FA model;

(h) when the constructed FA model matches a previously defined FA model, storing the identity of the constructed FA model as the final FA model along with the keywords; and

(i) when the constructed FA model does not match a previously defined FA model, generating a new FA model corresponding to the constructed FA model, storing the new FA model, and storing the identity of the new FA model as the final FA model along with the keywords.

According to one aspect of the present invention, the method also included steps for

(j) determining whether the keywords generated in step (e) match user-selected keywords; and

(k) when a match is not detected, terminating the multimedia signal parsing method.

These and other objects, features and advantages according to the present invention are provided by a combination receiving a multimedia signal including a video shot sequence, an audio signal and text information for performing story segmentation on the multimedia signal to generate discrete stories, each of which has associated therewith a final finite automaton (FA) model and keywords, at least one of which is associated with a respective node of the FA model. Advantageously, the combination includes:

a first device for analyzing the video portion of the received multimedia signal to identify keyframes therein to thereby generate identified keyframes;

a second device for comparing the identified keyframes within the video shot sequence with predetermined FA characteristics to identify a pattern of appearance within the video shot sequence;

a third device constructing a finite automaton (FA) model describing the appearance of the video shot sequence to thereby generate a constructed FA model;

a fourth device for coupling neighboring video shots or similar shots with the identified keyframes when the neighboring video shots are apparently related to a story represented by the identified keyframes;

a fifth device for extracting the keywords from the text information and storing the keywords at locations associated with each node of the constructed FA model;

a sixth device for analyzing and segmenting the audio signal in the multimedia signal into identified speaker segments, music segments, and silent segments

a seventh device for attaching the identified speaker segments, music segments, and silent segments to the constructed FA model;

an eighth device for storing the identity of the constructed FA model as the final FA model along with the keywords when the constructed FA model matches a previously defined FA model; and

a ninth device for generating a new FA model corresponding to the constructed FA model, for storing the new FA model, and for storing the identity of the new FA model as the final FA model along with the keywords when the constructed FA model does not match a previously defined FA model.